AMERICAN
STATISTICAL ASSOCIATION.

NEW SERIES, No. 92. DECEMBER, 1910.

THE CORRELATION OF ECONOMIC STATISTICS.

By Warren M. Persons, *Assistant Professor of Economics, Dartmouth College.*

The cause and effect relation existing between economic events is especially difficult to ascertain because of the presence of innumerable variable elements. In solving his problems the economist can not, like the physicist or chemist, eliminate all causes except one and then by experiment determine the effect of that one. Causes must be dealt with *en masse*. Since any effect is the result of many combined causes the economist is never sure that a given effect will follow a given cause. In stating an economic law he always has to postulate "other things remaining the same," with, perhaps, little appreciation of what the other things may be. It is rarely, if ever, possible for the economist to state more than "such and such a cause *tends* to produce such and such an effect." Events can only be stated to be more or less probable. He is dealing mainly, therefore, with correlation and not with simple causation.

The problems of economics are similar to certain problems of biology, such as the effect of environment and heredity upon the individual. In dealing with the question of heredity Karl Pearson says:* "Taking our stand then on the observed fact that a knowledge neither of parents nor of the whole ancestry will enable us to predict with certainty in a variety of important cases the character of the individual offspring we ask: What is the correct method of dealing with the problem of heredity in such cases? The causes A, B, C, D, E . . . which we have as yet succeeded in isolating and defining are not always followed by the effect X, but by any one of the effects U, V, W, X, Y, Z. We are therefore not dealing with causation but correlation, and there is therefore only one method of procedure possible; we must collect statistics of the frequency with which U, V, W, X, Y, Z, respectively, follow on A, B, C, D, E . . . From these statistics we know the most *probable* result of the causes A, B, C, D, E and the frequency of each deviation from this most probable result. The recognition that in the existing state of our knowledge the true method of approaching the problem of heredity is from the statistical side, and that the most that we can hope at present to do is to give the *probable* character of the offspring of a given ancestry, is one of the great services of Francis Galton to biometry."

Just as the biologists cannot predict a man's height or color of eyes or temper or combativeness by knowing those qualities in his ancestors, so economists cannot predict that a definite call rate in Wall Street will go with a given percentage of reserves to deposits in New York banks or that a given supply of wheat will result in a definite price per bushel. But, on the other hand, just as it has been observed that there *is* a relation existing between a man's stature and the stature of his ancestors, so it has been observed that a relation *does* exist between bank reserves and call rates and between supply of wheat and its price per bushel.

In order to deal in a satisfactory way with such questions as those given above it is necessary to accumulate statistics of the supposedly related phenomena. In order to have those

*The Law of Ancestral Heredity, Biometrika, Vol. II, p. 215.

statistics indicate anything it is necessary to obtain a method of measuring the extent of correlation between the phenomena.

The commonly used method of measuring the amount of correlation between any two series of economic statistics is to represent the two series graphically upon the same sheet of cross-section paper and then compare the fluctuations of one series with those of the other. The quantity theory of prices has been tested in this way by Dr. E. W. Kemmerer.* Dr. Kemmerer builds up the following price equation:

$$P_s = \frac{MR + CR_c}{NE + N_cE_c} \dagger$$

in which:

P_s = the average price (weighted by the total flows) of all commodities sold for money and deposit currency during a unit of time.

M = the total currency in circulation during the unit of time.

R = the average number of times each unit of currency changes hands during the unit of time.

} = the flow of currency

NE = the flow of goods exchanged for currency.

C = the volume of deposit currency exchanged for goods.

R_c = the average rate of turnover of such deposit currency.

} = flow of deposit currency.

N_cE_c = the flow of goods exchanged for deposit currency.

Dr. Kemmerer then attempts to find the answer that facts give to the following questions:

1. Do the bank reserves vary directly with the money supply?

2. Does the proportion of bank reserves to check circulation vary directly with the degree of business distrust existing in the country?

*Money and Credit Instruments in their Relation to General Prices, Henry Holt and Co., 1907.

†See Quarterly Journal of Economics, Feb., 1908, p. 274, for derivation of equation. In the current article, I quote from the review of Doctor Kemmerer's book.

3. Is "a relative increase in the circulating media accompanied by a corresponding and proportionate increase in general prices and a relative decrease in the circulating media, by a corresponding and proportionate decrease in general prices," or, in the language of the formula, is

$$P_s = \frac{MR + CR_c}{NE + N_c E_c} \quad *$$

borne out by the facts?

All of the questions to be tested by the statistics collected are questions of correlation. Dr. Kemmerer makes the tests graphically, as has been stated, by comparing the fluctuations of the two curves based upon the pair of series of statistics being considered. The charts presented by Dr. Kemmerer from which his conclusions are drawn are given below.

In the case of the correlation of bank reserves and money in circulation, inclusive of bank reserves, Dr. Kemmerer concludes, "There can be no question but that when due allowance is made for fluctuations in business confidence, the evidence of Chart I strongly supports the contention that there exists a close relationship between the amount of money in circulation and the amount of the country's bank reserves."† In the case of the correlation of business distrust and the ratio of bank reserves to check circulation the conclusion is, "the chart substantiates the contention . . . that the ratio of check circulation to bank reserves is a function of business

*Money and Prices, p. 139.

Dr. Kemmerer uses statistics of the United States for the period 1879–1904 to make his inductive tests. The statistics of total bank reserves he obtains from the Report of the Comptroller of the Currency. For the amount of money in each year (M) he takes the average of the total money in circulation at the beginning and end of that fiscal year as given in the Statistical Abstracts. The check circulation for each year he obtains by multiplying the total bank clearings in each year by $\frac{100}{38}$. This ratio is the ratio between the estimated total check circulation for 1896 and the bank clearings for that year. The rapidity of circulation of 47 per year he derives by dividing the estimated total money transactions in 1896 by the money circulation of that year. The figures for the growth of business he finds by taking the simple average of index numbers of fifteen different series of statistics taken as representing the industrial activity of the year considered. The index numbers of business distrust are the simple averages of the corresponding indices for the proportion of concerns failing, and the average liabilities of concerns failing. "The general index figures of prices and wages were computed by combining in a weighted average the index figures for the prices of railroad securities (Commons), the index figures for the prices of wholesale commodities (Commons), and the index figures for wages (Department of Labor tables for twenty-five occupations)."

†*Ibid.*, p. 143.

CHART I.

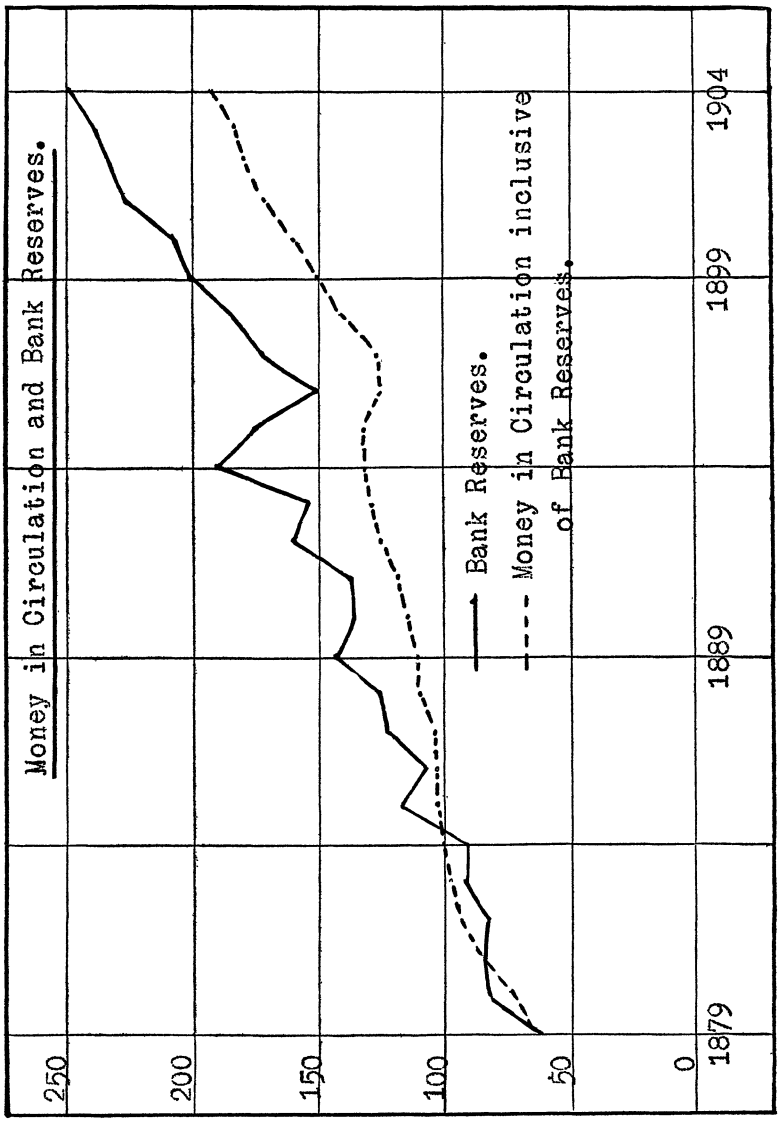

Money in Circulation and Bank Reserves.
250
200
150
100
50
0
1879
1889
1899
1904
Bank Reserves.
Money in Circulation inclusive of Bank Reserves.

CHART II.

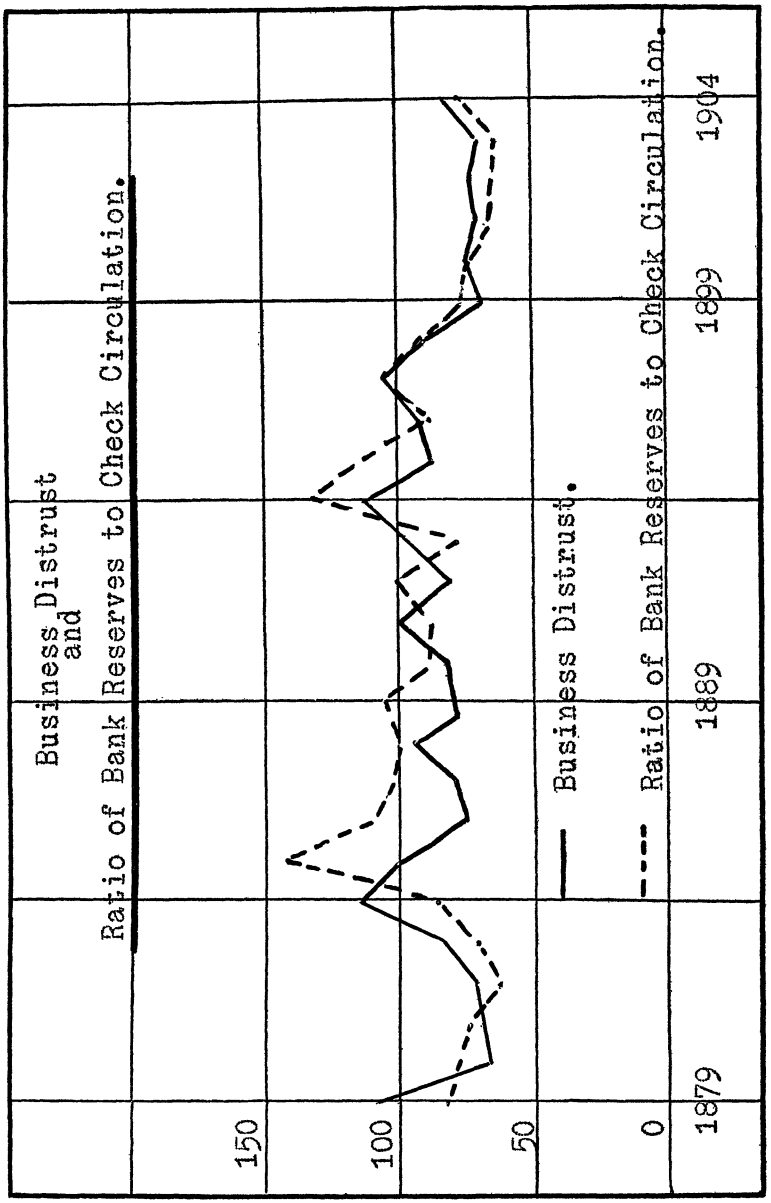
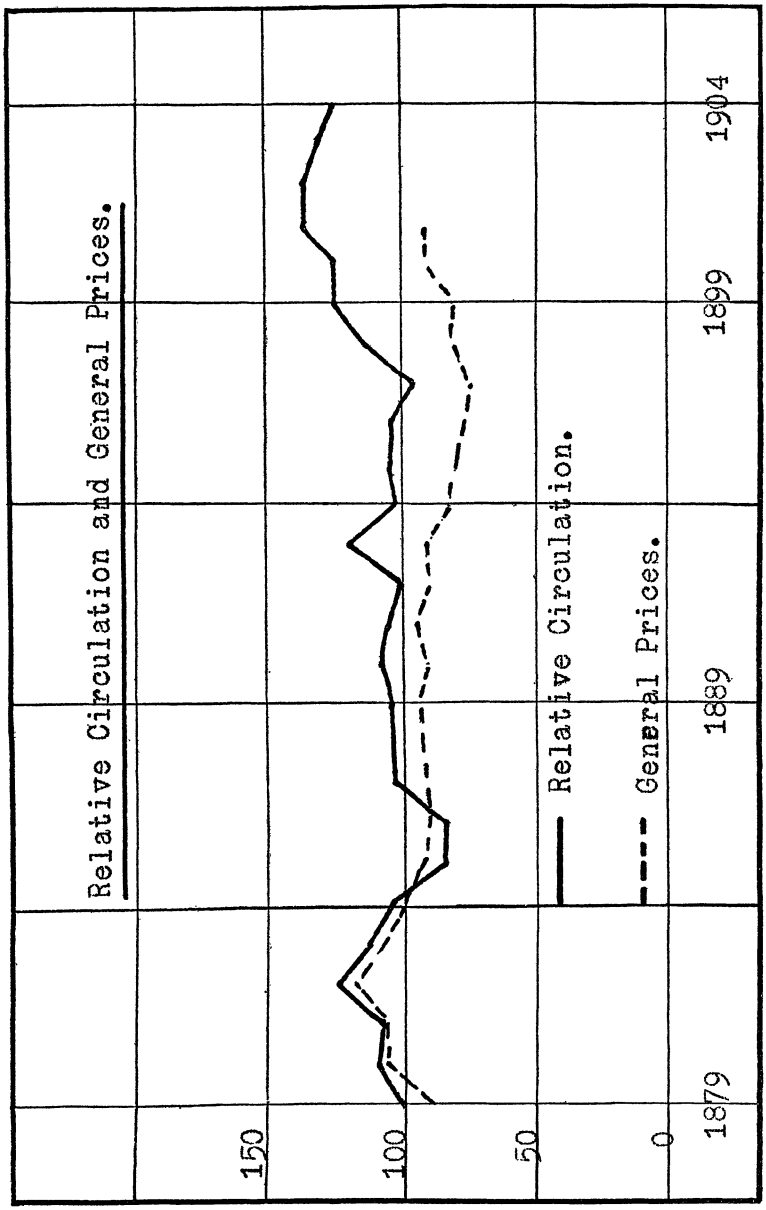

Business Distrust and Ratio of Bank Reserves to Check Circulation.
150
100
50
0
1879
1889
1899
1904
Business Distrust.
Ratio of Bank Reserves to Check Circulation.

CHART III.


Relative Circulation and General Prices.
150
100
50
0
1879
1889
1899
1904
Relative Circulation.
General Prices.

confidence . . ."* The final test of the quantity theory is the amount of correlation between the figures for the right and left-hand sides of the equation $P_s = \frac{MR + CR_c}{NE + N_cE_c}$. Upon examination of the curves plotted from the two series of statistics representing general prices and relative circulation (the left and right-hand sides, respectively, of the price equation) Dr. Kemmerer concludes, "The general movement of the two curves taken as a whole is the same, while the individual variations from year to year exhibit a striking similarity."†

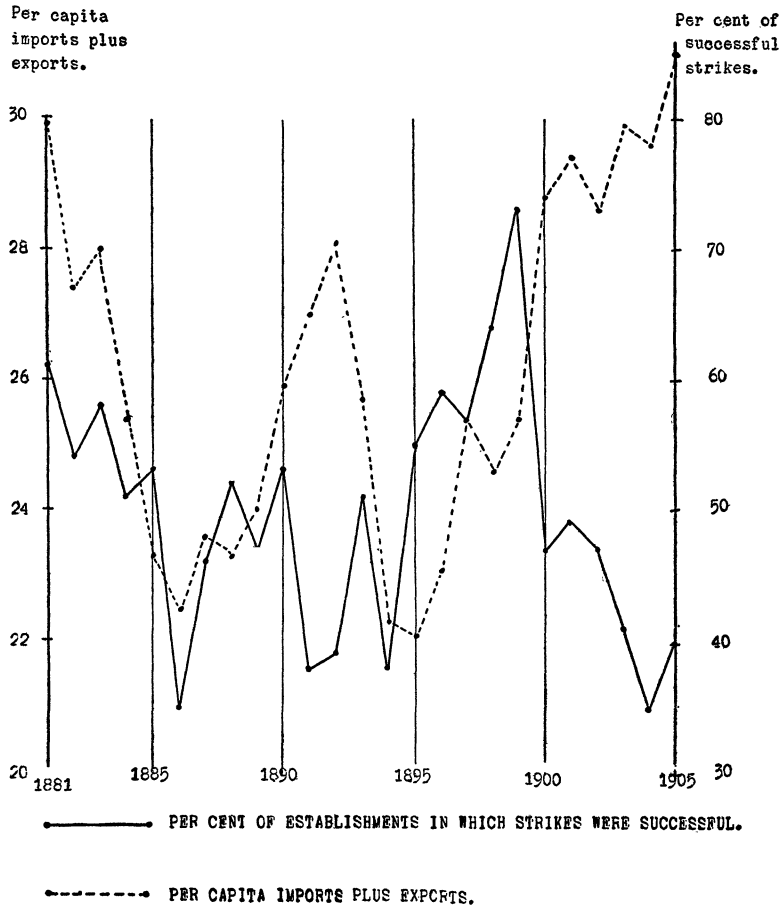
The graphic method of comparing fluctuations is well enough as a preliminary, *but does it enable anyone to tell anything of the extent of the correlation between the series of figures being considered?* Is Dr. Kemmerer warranted in deducing his conclusions from observation of the charts? It seems to the writer that one opposing the quantity theory might draw opposite conclusions with as much (or as little) reason. *The charts do not answer the questions proposed.* The painstaking collection of statistics to test correlation is useless if there be no more reliable method to measure correlation. A numerical measure of the correlation must be found if we wish to determine the *extent* to which the fluctuations of one series synchronize with the fluctuations of another series.

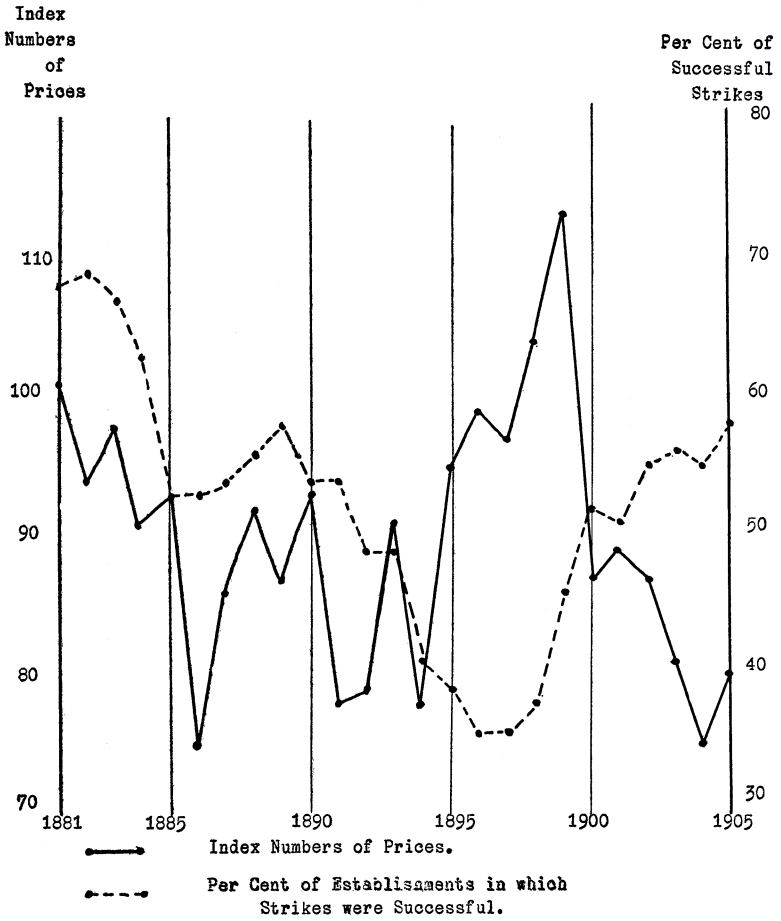
A second illustration of a conclusion based upon graphic representation is that of Ira Cross in his study of strike statistics.‡ He says, upon consideration of data taken from the Twenty-first Annual Report of the United States Bureau of Labor, "the percentage of successful strikes decreases during periods of business prosperity and increases during 'hard times.' " In the accompanying charts the per cent. of establishments in which strikes were successful is plotted, first, with the per capita exports and imports and second, with index numbers of wholesale prices. The foreign trade and the price statistics are taken as indicative of the activity of business, as indices of prosperity.

**Ibid.*, p. 146.

†*Ibid.*, p. 147.

‡Quarterly Publications of the American Statistical Association, June, 1908, p. 168.


Per capita imports plus exports.
Per cent of successful strikes.
30
28
26
24
22
20
80
70
60
50
40
30
1881
1885
1890
1895
1900
1905
PER CENT OF ESTABLISHMENTS IN WHICH STRIKES WERE SUCCESSFUL.
PER CAPITA IMPORTS PLUS EXPORTS.


Index Numbers of Prices
Per Cent of Successful Strikes
110
100
90
80
70
80
70
60
50
40
30
1881
1885
1890
1895
1900
1905
Index Numbers of Prices.
Per Cent of Establisaments in which Strikes were Successful.

A third illustration of a conclusion relating to correlation is taken from the *London Statist* of April 4, 1908, where the proposition is made that, "When commodities advance prices of Stock Exchange securities recede; when commodities recede Stock Exchange securities advance." The proposition is supported by reference to the following chart showing the yearly average price of consols and Sauerbeck's index numbers of prices.

The foregoing illustrations show the need by economists of a quantitative measure of correlation. Such a measure has been widely used in biological statistics and used to a limited extent in economic statistics.* G. U. Yule has used the measure in his study of "Pauperism;"† R. H. Hooker has used it in his "Correlation of the Weather and Crops;"‡ J. P. Norton applied it in his study of the "New York Money Market." This measure, the coefficient of correlation, will be computed for the data upon which the conclusions quoted above are based. The formula for the coefficient of correlation is

$$r = \frac{\Sigma xy}{n\sigma_1\sigma_2};$$

where:

x = deviation from arithmetic mean = $X - M_1$
y = deviation from arithmetic mean = $Y - M_2$
σ_1 = standard deviation of X series
σ_2 = standard deviation of Y series
n = number of items.

The coefficient of correlation "serves as a measure of any statement involving two qualifying adjectives, which can be measured numerically, such as 'tall men have tall sons,' 'wet springs bring dry summers,' 'short hours go with high wages.' " § It is not the purpose in what follows to go through the mathematical derivation of the coefficient of correlation, but to test the formula empirically in order to ascertain how it actually varies for given series of statistics and to point out some of its features.

However, it should be noted at this point that the coefficient of correlation is not empirical but was derived by *a priori* reasoning. It was found by assuming that a large number of independent causes operate upon each of the two

*See the very interesting paper by G. U. Yule on the Applications of the Method of Correlation to Social and Economic Statistics. (Journal of the Royal Statistical Society, December, 1909.) This paper gives a history of the application of the method of correlation with a bibliography of works in which it has been applied to economic and social statistics.

†Journal of the Royal Statistical Society, 1899, Vol. 62, pp. 249–286.

‡Journal of the Royal Statistical Society, March, 1907, p. 1.

§Bowley, A. L., Elements of Statistics, p. 320.

series X and Y, producing normal distributions in both cases. Upon the assumption that the set of causes operating upon the series X is *not independent* of the set of causes operating upon the series Y the value $r = \frac{\Sigma xy}{n\sigma_1\sigma_2}$ is obtained. This value becomes zero when the operating causes are absolutely independent. Hence the value of r was taken as a measure of correlation.* In what follows *no assumption concerning the type of distribution of the X and Y series will be made.*

Some appreciation of the meaning of the coefficient of correlation can be obtained by the consideration of a few simple applications. Suppose that we consider the two series of measurements:

X = 1, 2, 3, 4, 5 $\quad M_1 = 3$
Y = 6, 8, 10, 12, 14 $\quad M_2 = 10$

Deviations.		Square of Deviation.		Product of Deviations.	
x	y	x^2	y^2	xy	
−2	−4	4	16	8	$\sigma_1 = \sqrt{2}$
−1	−2	1	4	2	$\sigma_2 = 2\sqrt{2}$
0	0	0	0	0	$r = \frac{20}{5\sqrt{2}\cdot 2\sqrt{2}} = 1$
+1	+2	1	4	2	
+2	+4	4	16	8	

In the above illustration the numbers were chosen so that for an increase of 1 unit in the X series there is an increase of 2 units in the Y series. Thus the correlation is perfect and r equals +1. If the Y series had been 14, 12, 10, 8, 6 (the X series remaining the same) the value of r would have been −1. Thus −1 stands for perfect *negative* correlation, an increase in one series corresponding to a decrease in the other. It should also be noted in this connection that the coefficient of correlation (r) cannot be less than −1 nor more than +1.†

*See Yule, Journal of the Royal Statistical Society, Vol. 60, p. 839; Bowley, Elements of Statistics, p. 316; Elderton, Frequency Curves and Correlation, p. 106.

†Proof of this statement may be conveniently found in Bowley, Elements of Statistics, p. 319.

The above illustration suggests the question, "Will a linear relationship between X and Y *always* give perfect correlation?"

Assume the linear relationship

$$Y = aX + b$$

Since $y = Y - M_2$ and $x = X - M_2$

$$M_2 + y = a(x + M_1) + b \text{ or } y = ax$$

(since $b - aM_1 - M_2 = 0$)

$$\text{and } r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} = \frac{\Sigma ax^2}{\sqrt{\Sigma x^2 \cdot \Sigma a^2 x^2}} = \frac{a\Sigma x^2}{\sqrt{a^2(\Sigma x^2)^2}} = \pm 1$$

(The sign of r depends upon the sign of a.)

Therefore a linear relationship between two variables will give a correlation coefficient of $+1$ or -1 depending upon whether large values of one occur with large values of the other or large values of one occur with small values of the other.

The converse of the above proposition is likewise true, *i. e.*, if the coefficient of correlation (r) equals 1 then the relationship between the X and Y series is linear.

Assume $r = 1$

then $(\Sigma xy)^2 - \Sigma x^2 \cdot \Sigma y^2 = 0$

Letting $x_1 = \lambda_1 y_1$, $x_2 = \lambda_2 y_2$ . . . $x_n = \lambda_n y_n$ the above expression becomes

$$y_1^2 y_2^2 (\lambda_1 - \lambda_2)^2 + y_1^2 y_3^2 (\lambda_1 - \lambda_3)^2 + \ldots + y_r^2 y_s^2 (\lambda_r - \lambda_s)^2 + \ldots = 0$$

The only way in which this expression can equal zero is by having

$$\lambda_1 = \lambda_2 = \lambda_3 = \ldots = \lambda_n$$

and it follows that

$$x_1 = \lambda_1 y_1, \; x_2 = \lambda_1 y_2 \ldots x_n = \lambda_1 y_n$$

or

$$x = \lambda_1 y$$

which denotes a linear relationship between X and Y.

That any relation other than a linear one will not lead to r=1 is illustrated by the following:

Let $Y = X^2$

X = 1, 2, 3, 4, 5,	$M_1 = 3$
Y = 1, 4, 9, 16, 25,	$M_2 = 11$

x	y	x^2	y^2	xy	
−2	−10	4	100	20	
−1	− 7	1	49	7	$\sigma_1 = 1.41$
0	− 2	0	4	0	$\sigma_2 = 8.65$
+1	+ 5	1	25	5	$r = 0.981$
+2	+14	4	196	28	
Total		10	374	60	

Although the two series increase regularly, so that deviations of like signs always correspond, yet the correlation is not perfect *because a linear relation does not exist between X and Y.*

If the number of items in each series be increased to 11 and the Y items remain squares of the X's the value of r will be 0.974.

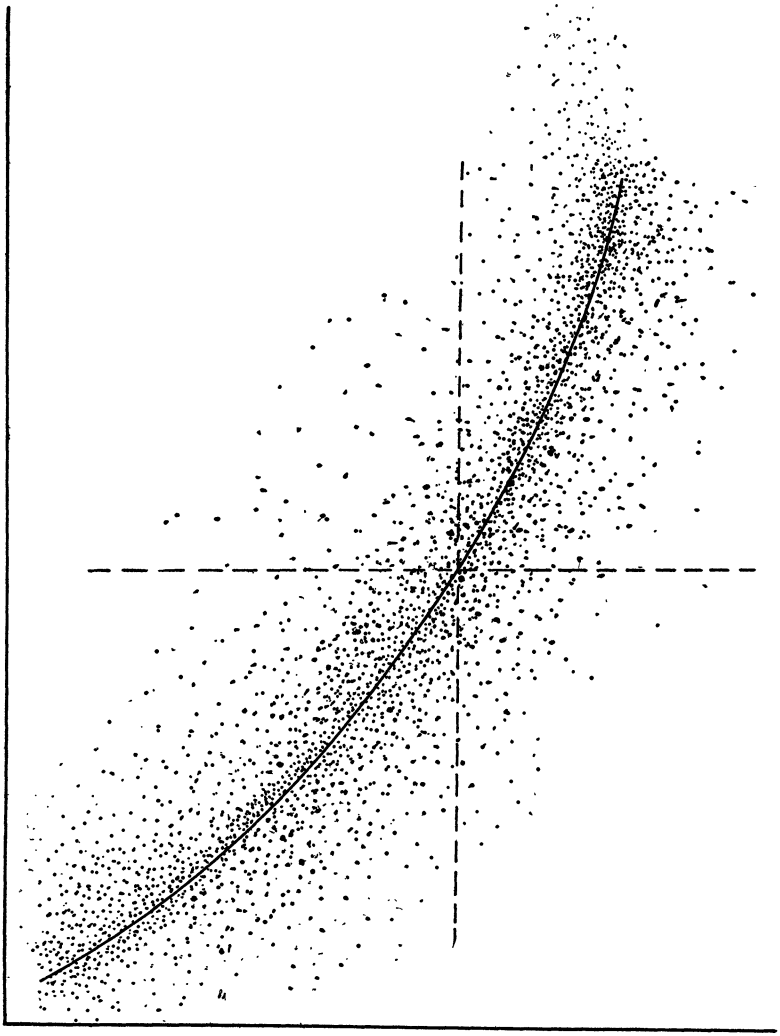
If there be no law connecting the X and Y series the products of the deviations (xy) are as apt to be negative as positive. The expression Σxy will therefore tend to approach zero. With a very large number of measurements the correlation coefficient will approximate zero.

From the condition of no relationship to the condition of a linear relationship existing between the pair of series of measurements the correlation coefficient varies from 0 to ± 1.

Suppose that we are investigating the relation existing between two series of measurements X and Y. Let points be plotted on cross-section paper whose coördinates are corresponding measurements X_i and Y_i. If there be a relationship existing between the two series, the points thus located will not lie chaotically all over the plane, but they will range themselves about some curve or locus. This curve, which has been called the *curve of regression,* is illustrated in the accompanying diagram (p. 302).* The straight line best fitting the points is called the line of regression.†

*Distributions similar to that of the diagram may be found in Natural Selection in "Helix Arbicstorum," A. P. di Cesuola, Biometrika, Vol. V, Part IV, p. 392, and Anthropometry of Scottish Insane, J. F. Tocher, Biometrika, Vol. V, Part III.

†So named by Francis Galton from the fact that if any group of men be picked out for an exceptional characteristic (say height) any other characteristic of those men correlated with height will regress toward the normal, or in other words the second characteristic will be less abnormal than the first.



For example suppose we consider the two series of index numbers for the period 1879–1904 inclusive, representing (1) money in circulation in the United States inclusive of bank reserves, and (2) bank reserves.* Let points be located with abscissas proportionate to the money in circulation and with ordinates proportionate to the bank reserves of the same year. The chart on the next page shows that these points lie near a straight line, the line of regression.

The coefficient of correlation (r) is a measure of the closeness of the grouping of the points about this line of regression. If the points should all range themselves on a line then r would equal $+1$ or -1 depending upon whether, looking left to right, the line sloped upward or downward.

We will now derive the equation of the line of regression. Let X and Y be associated measurements and x and y be associated deviations from the respective arithmetic means. A *linear* relation between the measurements is of the form

$$Y = a_1X + b_1$$

The relation between the deviations will be of form

$$y = a_1x \text{ or } y - a_1x = 0$$

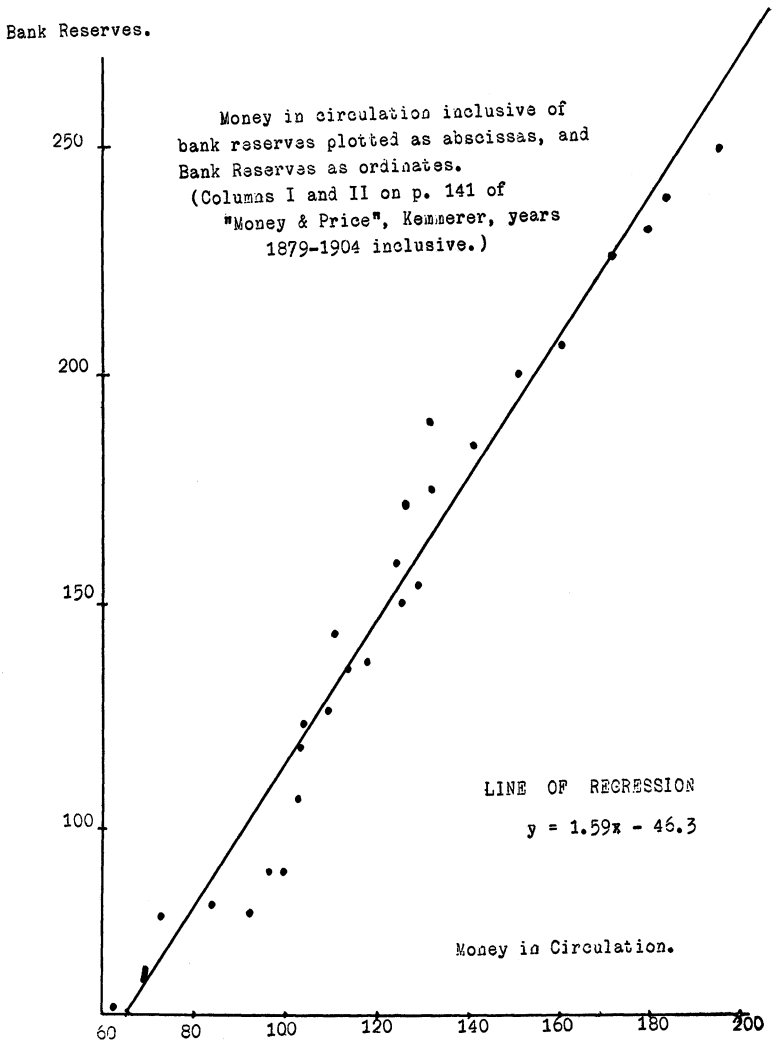
Since all of the points are not located exactly upon a straight line the substitution of the values $x_1\ y_1$, $x_2\ y_2$, etc. in the equations will give residues v_1, v_2, etc. as follows:

$$y_1 - a_1x_1 = v_1$$
$$y_2 - a_1x_2 = v_2$$
$$y_n - a_1\ x_n = v_n$$

The values $\frac{v_1}{\sqrt{1+a_1^2}}, \frac{v_2}{\sqrt{1+a_1^2}}, \ldots \frac{v_n}{\sqrt{1+a_1^2}}$ equal the distances of the various points to the straight line $y = a_1x$.

The equation of a line such that the sum of the squares of the distances from the given points is a minimum will now be found. In other words that value of a_1 will be taken which makes $v_1^2+v_2^2+ \ldots v_n^2 = a$ minimum. To find the value of a_1, for which $(y_1-a_1x_1)^2+(y_2-a_1x_2)^2+ \ldots +(y_n-a_1x_n)^2$ will be a minimum, differentiate with respect to a_1 and obtain $-2x_1(y_1-a_1x_1)-2x_2(y_2-a_1x_2) \ldots -2x_n(y_n-a_1x_n)$. In order that the original function be a minimum, this derivative must equal zero. We will then have

*Kemmerer, Money and Prices, p. 141.


Bank Reserves.
Money in circulation inclusive of bank reserves plotted as abscissas, and Bank Reserves as ordinates.
(Columns I and II on p. 141 of "Money & Price", Kemmerer, years 1879-1904 inclusive.)
250
200
150
100
LINE OF REGRESSION
y = 1.59x - 46.3
Money in Circulation.
60
80
100
120
140
160
180
200

$(x_1y_1 - a_1x_1^2) + (x_2y_2 - a_1x_2^2) \quad . \quad . \quad . \quad . \quad + (x_ny_n - a_1x_n^2) = 0$, or $\Sigma xy - a_1\Sigma x^2 = 0$

$$a_1 = \frac{\Sigma xy}{\Sigma x^2}$$

Similarly if $x = a_2y$, then $\Sigma xy - a_2\Sigma y^2 = 0$ will give the value of a_2 for which the sum of the squares of the distances of the given points to the straight line $X = a_2Y + b_2$ is a minimum, or $a_2 = \frac{\Sigma xy}{\Sigma y^2}$

Let $r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} = \frac{\Sigma xy}{n\sigma_1\sigma_2}$ and $\Sigma x^2 = n\sigma_1^2$ $\Sigma y^2 = n\sigma_2^2$.

The equations between the deviations are:

$$y = r\frac{\sigma_2}{\sigma_1}x$$

$$x = r\frac{\sigma_1}{\sigma_2}y$$

It may seem that the two equations just given are inconsistent. But it must be remembered that these equations do not give the relationship existing between *any* corresponding pair of deviations unless all of the points lie exactly on a straight line and there be perfect correlation. For all cases of imperfect correlation a *given* deviation x will occur with several different deviations y (if we have a large number of measurements). If these deviations y are distributed according to the normal law of distribution then the given value x substituted in the first equation will give the mean of the deviations occurring with the deviation x and if a given value y be substituted in the second equation the value of x resulting will be the mean of the deviations of the associated characteristics.

Since $y = Y - M_2$ and $x = X - M_1$

$$Y = M_2 + r\frac{\sigma_2}{\sigma_1}(X - M_1)$$

and $$X = M_1 + r\frac{\sigma_1}{\sigma_2}(Y - M_2)$$

The coefficients $r\frac{\sigma_2}{\sigma_1}$ and $r\frac{\sigma_1}{\sigma_2}$ are called the coefficients of regression of Y upon X and of X upon Y respectively. The

first coefficient ($r\frac{\sigma_2}{\sigma_1}$) and the reciprocal of the second ($\frac{\sigma_2}{r\sigma_1}$) are the *slopes* of the lines of regression. If X and Y be measured in terms of their respective standard deviations as units the slopes of the lines of regression will be r and $\frac{1}{r}$. *In other words, the slope of the line of regression of Y upon X, each series being measured in terms of its standard deviation, is equal to the coefficient of correlation for the two series. For perfect positive correlation the line would make an angle of 45° with the X axis for perfect negative correlation the line would make an angle of 135° with the x axis, and for no correlation the line would be parallel to the x axis.*

In the preceding we have fitted a straight line to the points. In some cases the points may arrange themselves along some locus such as could be represented by $y=a+bx+cx^2+ \ . \ . \ . \ . \ .$ or $y=a+bc^x$ or any other empirical function. The work in fitting such a curve to the points would be great and the advantages are not sufficient to compensate for the additional work.*

The following table gives the correlation coefficients for the various pairs of series of statistics for which Dr. Kemmerer has attempted to show correlation by means of charts:

Statistics of	Period Covered.	Coefficient of Correlation ± Probable Error.
Money in circulation inclusive of bank reserves / Bank Reserves	1879–1904	+0.98 ± .006
Business distrust / Ratio of bank reserves to check circulation	1879–1904	+0.53 ± .095
Business distrust / Ratio of bank reserve to check circulation	1879–1903 / 1880–1904	+0.72 ± .064
Relative circulation / General Prices	1879–1901	+0.23 ± 0.13

*Karl Pearson in an article "On the General Theory of Skew Correlation and Non-linear Regression" (published by Dulau and Co. in 1905 as one of the Drapers' Company Research Memoirs) discusses the cases in which the law connecting the two series of statistics is not linear. He says, "In the great bulk of biometrical and economical enquiries, however, the regression does not diverge very markedly from the linear form. In the cases of non-linear regression that I have hitherto had to deal with, I find that parabolae of the second or third order will suffice as a rule to describe the deviations from linearity" (p. 21). Equations of parabolae of the second and third order are of the form

$y=a_0+a_1x+a_2x^2$

and $y=a_0+a_1x+a_2x^2+a_3x^3$

The correlation coefficients show that there is a very great difference in the degree of correlation of different pairs of series of statistics. The full significance of the "probable error," which is used as a measure of unreliability of any determination, cannot be developed at this point.* It is sufficient to note that, "When r is not greater than its probable error we have no evidence that there is any correlation, for the observed phenomena might easily arise from totally unconnected causes; but, when r is greater than, say, six times its probable error, we may be practically certain that the phenomena are not independent of each other, for the chance that the observed results would be obtained from unconnected causes is practically zero."†

The high degree of correlation (+0.98) between money in circulation inclusive of bank reserves and bank reserves is due to the tendency of the two items to vary together during the long time period and not due to correspondence of minor fluctuations. The reasons for the great increase of money in circulation in the United States during the period 1879–1904 are the great increase of population and the industrial expansion. Likewise the number of banks increased in order to serve the increased population and this meant an increase of total reserves. It is self-evident that the long time tendency of the two series of statistics must be upward in a growing country. It seemed to me that the bank reserves during the 26 years, 1879–1904, would be as closely correlated with the *population* as with total circulation. The computation of the correlation coefficient between bank reserves and population gave +0.98. It is the variation upwards of both series during the entire period that causes the high coefficient.

The correlation coefficient between the index numbers of business distrust and the rates of bank reserves to check circulation for the same years is 0.53. When the index numbers of business distrust for one year are correlated with the ratio of bank reserves to check circulation the following year the coefficient is 0.72. As Dr. Kemmerer has suggested (but

*In the computation of the formulas for the probable error it is assumed that errors are distributed "normally."

†Bowley, Elements of Statistics, p. 320.

not verified), there is a closer correlation "when proper allowance is made for the time required for alterations in business confidence to exert their influence on bank reserves."* The lowest correlation (+0.23), that between relative circulation and general prices, is not high enough to warrant a conclusion that the items vary together. The smallness of the correlation indicated may have resulted either because the quantity theory is in error or because the statistics are not adequate to test the theory. Whatever may be the fact, the statistics and the method of measuring correlation presented by Dr. Kemmerer do not demonstrate that general prices move in sympathy with relative circulation.

Is the contention of Mr. Cross that "the percentage of successful strikes decreases during periods of business prosperity and increases during 'hard times'" supported by the statistics? The following table gives the correlation coefficients between pairs of series of statistics, one of the pair in each case being the per cent. of establishments in which strikes succeeded, and the other series being taken as indicative of business conditions:

Statistics of	Period	Coefficient of Correlation ± Probable Error
Per cent. of establishments in which strikes succeeded / Index numbers of wholesale prices from Aldrich Report and United States Labor Bureau†	1881–1905	−0.146 ± 0.132
Per cent. of successful strikes / Prices	1881–1904 1882–1905	−0.086 ± 0.134
Per cent. of successful strikes (year ending Dec. 31) / Per capita foreign trade (year ending June 30)	1881–1905	−0.178 ± 0.130
Per cent. of successful strikes / Index numbers of business distrust‡	1881–1904 1881–1904	−0.076 ± 0.134

The amount of correlation indicated in each case is small—considering the number of years taken, so small that no conclusion as to the connection between the two series can be

*Money and Prices, p. 146.

†Reduced to a continuous series.

‡Kemmerer's figures from Money and Prices, p. 141.

drawn. The correlation coefficient in the last instance, *i. e.*, between per cent. of successful strikes and business distrust, suggests an opposite conclusion to that indicated by the other coefficients and that of Mr. Cross. The analysis shows that the conclusion that there is negative correlation between *general* prosperity and per cent. of successful strikes is not warranted.

Finally, what is the degree of correlation between the prices of British Consols and Sauerbeck's index numbers of the prices of commodities? The chart on p. 297 indicates a greater degree of correlation (negative) between the *minor* fluctuations of the two series than shown by any of the pairs of series that we have considered. The coefficient of correlation based upon statistics for the 57 years from 1851 to 1907, inclusive, is -0.58 ± 0.06. A correlation coefficient of -0.58 based upon 57 pairs of items warrants the conclusion that the two series have inverse movements.

The relations between the *average* deviations, x and y, of the two series of statistics being considered are:*

$$y = -1.465x \text{ and } x = -0.2295x$$

The equations of regression are:

$$Y = 225.6 - 1.465\ X \text{ and } X = 19.439 - 0.2295\ Y$$

For certain pairs of time-series (corresponding items occur at same time) of measurements a correlation coefficient approximating zero may be obtained even though graphs of the statistics show that the up-and-down fluctuations occur together. This result will come about if the *long-time* variations show opposite tendencies, as, for instance, in the statistics of marriages and bank clearings in the United Kingdom. On the other hand, a *high* correlation coefficient may be obtained for two series having the same long-time tendency regardless of the non-correspondence of the short-time fluctuations. For example, the coefficient for the two series, population and bank reserves, came out to be 0.98. This high coeffi-

*If a value x_1 be substituted for the deviation x in the equation $y = r\frac{\sigma_2}{\sigma_1}x$ we ought to get an approximation to the average value of the *deviations* of the Y character, all of which characters are associated with the deviation x_1 of the X character. Likewise if y_1 be substituted in $x = r\frac{\sigma_1}{\sigma_2}y$ we ought to get an approximation to the average of the associated x deviations. The closeness of the approximation will increase as the form of distribution of one series of characters associated with a given value of the other character (called an array) approaches the normal curve of error.

cient comes from the fact that the long-time variation of both series is the same. Consequently, before it is legitimate to draw any conclusions as to the meaning of a lack of correlation, or amount of correlation between two series of measurements it is necessary to ascertain the periodic and the secular variations in the two series. This correlation coefficient mav be large through the correspondence of either secular or periodic variation, or both. It may be null because one variation covers up the other.

Three methods have been used for isolating the short-time variations of time-series of measurements. They will now be considered.

1. If upon plotting the two series being compared with time as abscissa and the measurements as ordinates, *periodic* variations appear at approximately equal intervals of time the curve may be "smoothed" and the secular variations may be eliminated as follows:*

(a) Ascertain the length of the wave by finding the number of time units between corresponding parts of the waves, *i. e.*, crest to crest, or hollow to hollow. Let l represent the number of time units found.

(b) Average groups of l consecutive measurements, placing the points, determined by these averages at the middle of each group of measurements. Take enough groups so that the points obtained will indicate the general tendency of the series.

(c) Draw a smooth curve through the points located by the process described in (b). This curve shows the secular tendency.

(d) Subtract (this can be done graphically on cross-section paper) the ordinates of the "smoothed" curve from those of the original curve in order to obtain the series of measurements of the periodic fluctuation. Let d stand for any one of these differences.

(e) The coefficients computed for corresponding ordinates of two smoothed curves, and for corresponding differences, d and d′, give measures of the secular and periodic correlation, respectively.

*Bowley, Elements of Statistics, pp. 176, *et seq.*

The method described above has been applied by Mr. R. H. Hooker in his paper "On the Correlation of the Marriage-Rate with Trade,"* and by Mr. G. U. Yule in his study of "Changes in Marriage and Birth-Rates in England and Wales during the Past Half Century."† The following table gives the correlation coefficients computed in the articles named for the *periodic* variations:

Series	Period	Deviations from	Coefficient of Correlation
Marriage rate..... Imports plus exports per capita.....	1861–1895	9 yr. means	+0.86
Marriage rate..... Amount of bank clearings per capita	1876–1895	9 yr. means	+0.47
Marriage rate..... Sauerbeck's index numbers of prices.	1865–1896	11 yr. means	+0.795
Marriage rate..... Hartley's index numbers of unemployment	1870–1895	11 yr. means	−0.873

The effect of using the deviations rather than the original series in computing the coefficient is shown by the comparison of the first correlation coefficient of +0.86, given above, with the correlation coefficient of +0.18, obtained for the same two series of *original* measurements for the same period, 1861–1895.

Using the deviation-method, Mr. Yule computed the correlation coefficients between *first*, the marriage rate of one year (m), and *second*, exports (e), imports (i), total trade (t), the price of wheat (w), and bank clearings (c) for the same year, and for each of several preceding years in order to answer the question, "does the maximum amount of correlation occur when corresponding items are of same year or when the marriage rate of one year is paired with the business item for a preceding year?"

*Journal of the Royal Statistical Society, September, 1901, Vol. 60.

†*Ibid.*, Vol. 69, pp. 88–132.

The following table gives the *maximum* values found:

Series Correlated	For period 1861–1895		For period 1876–1895	
	Business item precedes marriage-rate item by	Value of Coefficient	Business item precedes marriage-rate item by	Value of Coefficient
rme	½ year	+0.86	½ year	+0.90
rmi	½ year	+0.82	½ year	+0.86
rmt	½ year	+0.91	½ year	+0.92
rmw	.....	.....	0 year	+0.56
rmc	.....	.....	1 year	+0.92

Having found that the trade cycle affects marriage rate Mr. Yule asks the question, "Does the cycle affect the birth-rate, and how?* The relation between marriage-rate and birth-rate is shown by the following table:

Series	Period	Deviations from	(a) Correlated with (b) of	Correlation Coefficient ± P. E.
(a) Marriage-rate (b) Birth-rate	1850–1896	11 year means	1 yr. following 2 yrs. following 3 yrs. following	0.352±0.086 0.479±0.076 0.418±0.082

Mr. Yule says, "Fitting a parabola to the three values thus determined, a maximum correlation of about 0.482 must subsist between the birth-rate and the marriage-rate of 2.17 (two years and two months) previously."†

Further analysis leads Mr. Yule to the conclusion that birth-rate is independently (not through marriage-rate only) sensitive to short-time economic changes and that the birth-rate is lowered after a depression, not only because of a decrease in the number of marriages during such depression, but also to a decrease in fertility.

2. In case the statistics show a long-time tendency with no *regular periodic* fluctuation Mr. R. H. Hooker has suggested that the "differences between successive values of the two

**Ibid.*, p. 123.

†*Ibid.*, p. 123.

variables, instead of the differences from the arithmetic means"* be correlated. Put into mathematical symbols we have:

Letting $\left\{\begin{matrix} X_o, X_1 \ldots X_n \\ X'_o, X'_1 \ldots X'_n \end{matrix}\right\}$ represent two series of measurements,

and $\left\{\begin{matrix} d_1, d_2 \ldots d_n \\ d'_1, d'_2 \ldots d'_n \end{matrix}\right\}$ represent differences between any two consecutive measurements,

and $\left\{\begin{matrix} d_m \\ d'_m \end{matrix}\right\}$ represent the respective means of these differences,

then $d_m = \frac{X_n - X_o}{n} = \frac{\Sigma d}{n}$, and

$$d'_m = \frac{X'_n - X'_o}{n} = \frac{\Sigma d'}{n};$$

and the standard deviations of the differences are

$$\delta = \sqrt{\frac{\Sigma(d-d_m)^2}{n}}$$

$$\delta' = \sqrt{\frac{\Sigma(d'-d'_m)^2}{n}};$$

and the coefficient of correlation is

$$\rho = \frac{\Sigma(d-d_m)(d'-d'_m)}{n\delta'\delta} = \frac{\Sigma dd' - nd_m d'_m}{\sqrt{(\Sigma d^2 - nd_m^2)(\Sigma d'^2 - nd'^2_m)}}$$

Comparing this method of differences with the method described in (1) Mr. Hooker says, "Correlation of the deviations from an instantaneous average (or trend) may be adopted to test the similarity of more or less marked periodic influences, correlation of the difference between successive values will probably prove most useful where the similarity of the shorter rapid changes (with no apparent periodicity) are the subject of investigation, or where the normal level of one or both series of observations does not remain constant."† He finds that the ordinary correlation coefficient (r) for the price of corn in Iowa and total production in the United States for the period 1870–1899 is –0.28, while $\rho = -0.84$.

**Ibid.*, Vol. 68, p. 697.

†*Ibid.*, p. 703.

I have computed ρ for the statistics of corn production in the United States and the average farm price on December 1* for the period 1866–1906 and finds $\rho = -0.833 \pm 0.034$. Letting x represent the production *difference* in millions of bushels, and y represent the price *difference* in cents per bushel, the equations of regression are

$$y = -0.0256x + 1.132$$
$$x = -27.05y + 46.42$$

A graphic representation of the points whose abscissas and ordinates are the corresponding production and price differences, respectively, and the line of regression is given on page 315. The lack of correlation between the original pair of series is shown by the chart on page 316.

From the equations of regression such statements as the following can be made:†

(i) For no change in corn production there is an increase in price of 1.132 cents per bushel.

(ii) For an increase in production of 100 million bushels the price decreases 1.43 cents per bushel.

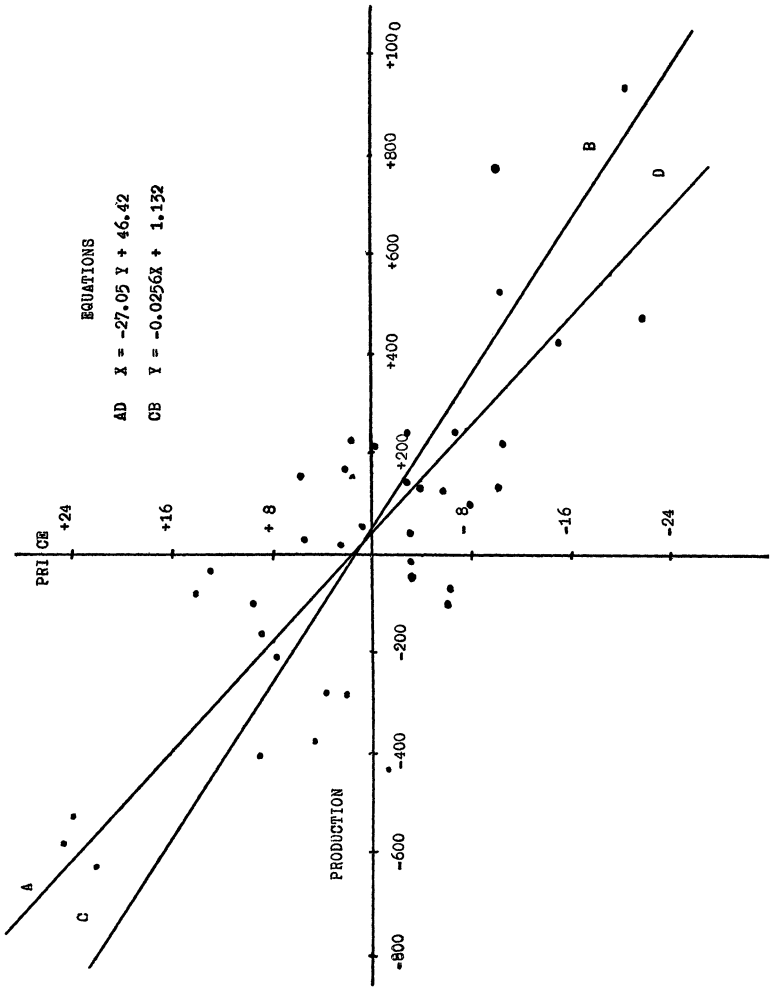
(iii) For a decrease in production of 100 million bushels the price increases 3.69 cents per bushel.

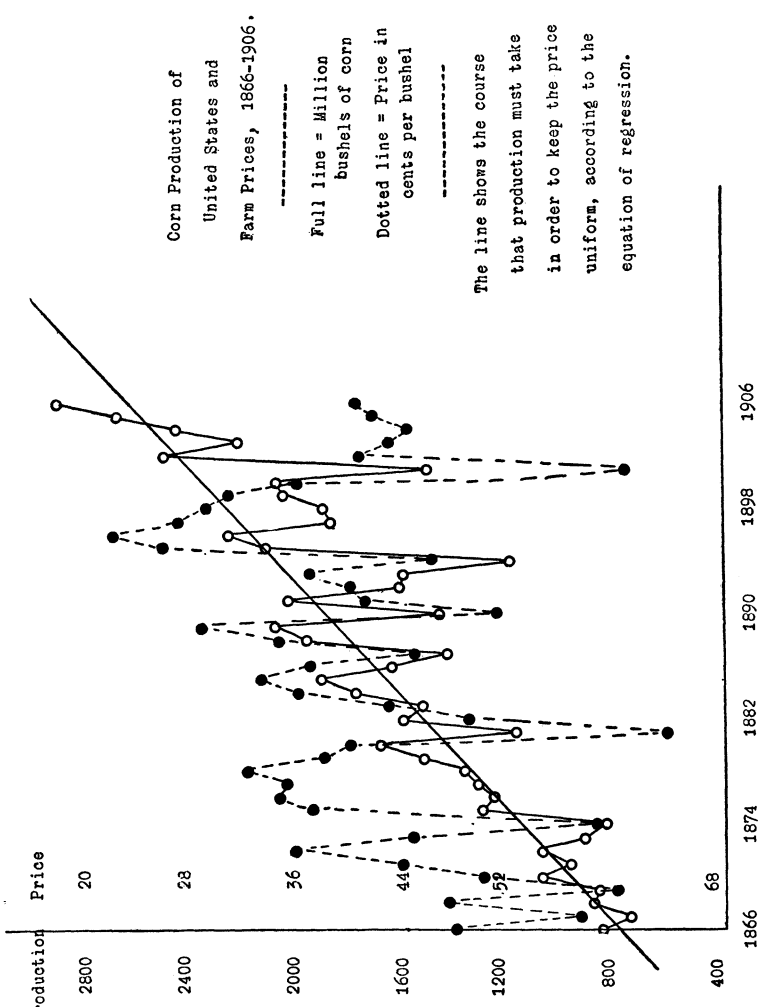
(iv) For a stationary price the production must increase 46 million bushels per year.

It seemed to me that if *percentage* changes in price and production were used instead of absolute changes a still closer correlation might result. The computation of ρ from such percentages, however, gave -0.794.

*Statistical Abstract of the United States, 1906, p. 543.

†The writer is making similar computations for wheat, oats, rye, barley, buckwheat, pig-iron, wool, bituminous coal and anthracite coal. It seems to me that the determination of the correlation between money and prices might be carried out by this method. After having allowed for the influence of changes in the supply of the various articles on those articles the influence of the change in the supply of money upon all the articles might be determined.


EQUATIONS
AD X = -27.05 Y + 46.42
CB Y = -0.0256X + 1.132
PRICE
PRODUCTION
+24
+16
+8
-8
-16
-24
+1000
+800
+600
+400
+200
-200
-400
-600
-800
A
B
C
D


Corn Production of
United States and
Farm Prices, 1866-1906.
Full line = Million
bushels of corn
Dotted line = Price in
cents per bushel
The line shows the course
that production must take
in order to keep the price
uniform, according to the
equation of regression.
Production
Price
2800
2400
2000
1600
1200
800
400
20
28
36
44
52
68
1866
1874
1882
1890
1898
1906

In the preceding illustrations the amount of correlation between the differences was greater than that between the original series. The method of differences has also been used by the writer for Kemmerer's statistics (considered on page 15 of this article) of (1) money in circulation, and (2) bank reserves for the period 1879–1904 with the result $\rho = +0.392$, whereas the value of r is 0.98. This shows that there is a lack of correspondence of the short-time variations in these two series.

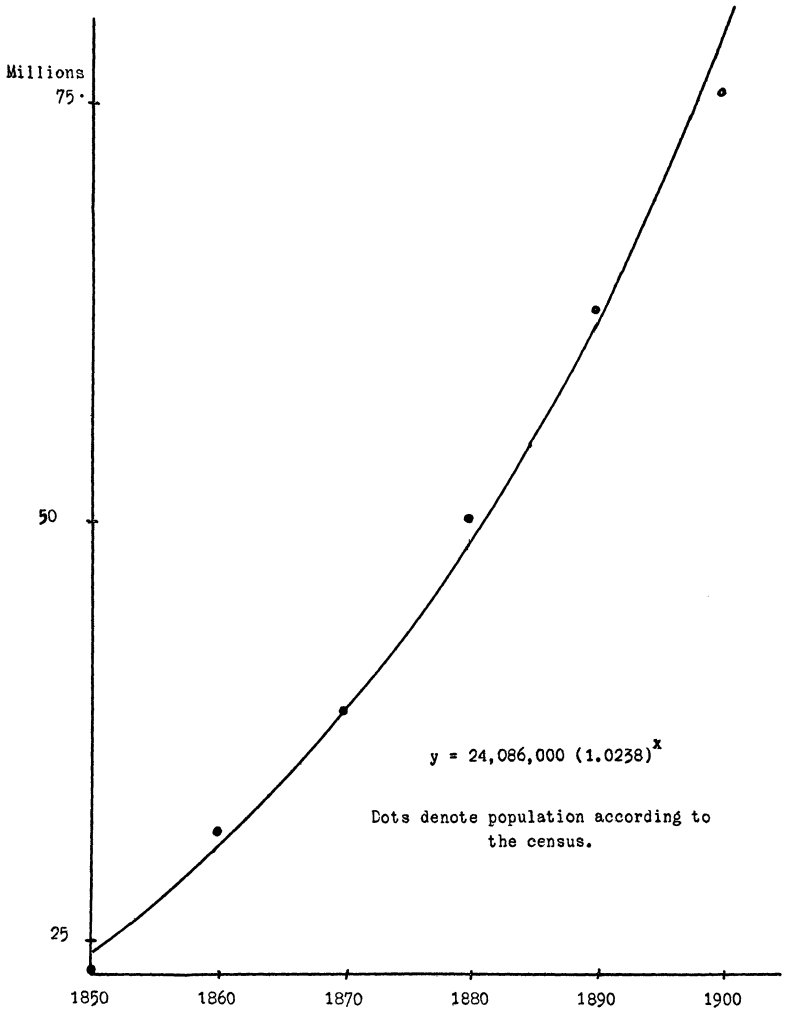
3. A third method of eliminating the long-time tendency and thus isolating the short-time fluctuations is to assume some curve, represented by an algebraic equation, which "fits" the statistics in question. The first step in the process is to select some curve, which, for *a priori* or other reasons is considered the best representation of the "growth element."* The second step is to fit the curve to the statistics; stated algebraically, to determine the constants in the equation of the curve by use of the actual data.† Finally the deviations of the original measurements from the smooth curve (called by Norton "the growth axis") are computed. The accuracy with which one law, the geometric, $y = bc^x$, describes the population of the United States, and consequently many things that depend upon population is shown by the following diagram. The full points are fixed by the actual population according to each of the censuses from 1850. The smooth line is the graph of the equation

$$y = 24{,}086{,}000\ (1.0238)^x,$$

which equation was determined from the actual population.

*J.P. Norton gives a table of interpolation forms from Steinhauser on p. 25 of the New York Money Market.

†The method of fitting curves to statistics constitutes a separate subject. An extensive use of mathematics is necessary in order to develop this subject fully.


Millions
75
50
25
1850
1860
1870
1880
1890
1900
y = 24,086,000 (1.0238)^x
Dots denote population according to the census.

Prof. J. P. Norton has applied the method here described to determine the correlation existing between percentage of reserves to deposits of New York Associated Banks and call rates.* Weekly statistics were taken for the period 1885–1900. The growth axes assumed were the geometric curve, $y = bc^x$, and the straight line $y = a$, respectively. (y = the measurement, x = time measured in weeks, while a, b and c are constants to be determined from the data.) The typical periodic fluctuations of *percentage* deviations of reserves and loans were also correlated by this method, using $y = bc^x$ as the growth function in both cases. The following table gives the correlation coefficients, ρ †

Series	ρ
Reserve deviations and discount rate	-0.37 ± 0.02
(a) Reserve and (b) Loan Periods Immediate	$+0.49 \pm 0.07$
(a) precedes (b) by one week	$+0.62 \pm 0.06$
(a) precedes (b) by two weeks	$+0.87 \pm 0.02$
(a) precedes (b) by three weeks	$+0.96 \pm 0.01$
(a) precedes (b) by four weeks	$+0.91 \pm 0.05$

The conclusion from this study is that "the loan period is really the shadow of the reserve period" . . . and apparently follows the latter by "an interval of approximately three weeks."‡

Up to this point the problem before us has been the measurement of the amount of correlation between two variables. This is the simplest case of the general problem of the measurement of the amount of correlation between one series of measurements, and a group of any number of series of measurements. The solution of the general problem leads to very complex relations,§ and it will not be taken up here. The case of three variables will be considered briefly.

*J. P. Norton, Statistical Studies in the New York Money Market.

†*Ibid.*, p. 96.

‡*Ibid.*, p. 94.

§Yule, G. U., on the Theory o Correlation, Journal of the Royal Statistical Society Vol. 60, p. 835.

Messrs. R. H. Hooker and G. U. Yule have considered the problem, To find the relation between the production of wheat in India during the period 1890–1904 (years ending March 31), the price of wheat (calendar years), and the exports of the subsequent twelve months, 1891–1905 (years ending March 31). The correlation of the annual differences according to the method described in (2) of page 312 gives the following results:

Series Correlated	Coefficient Correlation
1. Exports and Production	+0.77
2. Exports and Price	+0.86
3. Production and Price combined in the ratio 1: 1, and Exports	+0.90
4. Production and Price combined in the ratio 3: 1, and Exports	+0.81
5. Production and Price in the ratio 1: 3, and Exports	+0.58

The table indicates that exports depend upon production and price, and depend equally upon them.

Messrs. Hooker and Yule give the following general solution of the special problem just considered:

To find the maximum correlation coefficient between x_1 and x_2+bx_3 that results from considering b a variable, where x_1, x_2, and x_3 are the deviations of the series X_1, X_2, and X_3 from their respective arithmetic averages.

Let $x_2+bx_3=z$

then $\Sigma(x_1z)=\Sigma x_1x_2+\Sigma bx_1x_3=n(r_{12}\sigma_1\sigma_2+br_{13}\sigma_1\sigma_3)$

and $\Sigma z^2=n(\sigma_2^2+b^2\sigma_3^2+2br_{23}\sigma_2\sigma_3)$

Hence $$\sqrt{x_1z}=\frac{r_{12}\sigma_2+br_{13}\sigma_2}{\sqrt{\sigma_2^2+b^2\sigma_3^2+2br_{23}\sigma_2\sigma_3}}$$

To find the value of b for which this is a maximum, differentiate with respect to b and equate to zero; then

$$b=\frac{(r_{13}-r_{12}\cdot r_{23})\sigma_2}{(r_{12}-r_{13}\cdot r_{23})\sigma_3}$$

which gives the maximum value

$$\sqrt{x_1z}=\sqrt{\frac{r_{12}^2+r_{13}^2-2r_{12}r_{23}r_{31}}{1-r_{23}^2}}$$

Computing $\sqrt{x_1z}$ from the data of Indian production, price, and exports of wheat the value 0.905 is obtained.

Mr. G. U. Yule, in the paper already referred to,* has worked out the general solution of the problem of the correlation between three variables. In the course of the solution the problem just considered is solved incidentally. The argument is similar to that used in the case of two variables and so it will not be repeated here. A concrete notion of the results secured by Mr. Yule can be obtained from the following explanation taken from Mr. Hooker's article on the "Correlation of the Weather and the Crops."†

"I have in the first place formed the ordinary coefficient $r = \frac{\Sigma(xy)}{\sqrt{n\sigma_1\sigma_2}}$ between the crop and (a) rainfall, (b) accumulated temperature above 42°. But rainfall and temperature are themselves correlated; hence an apparent influence of, say, rainfall upon a crop may really be due to rainfall conditions being dependent upon temperature, or *vice versa*. Hence it seemed desirable to calculate the *partial* or *net* correlation coefficients, *i. e.* (following the notation given in Mr. Yule's paper of 1897).

$$\rho_{12} = \frac{r_{12} - r_{12}\cdot r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}, \quad \rho_{13} = \frac{r_{13} - r_{12}\cdot r_{23}}{\sqrt{(1-r_{23}^2)(1-r_{12}^2)}}$$

"This partial coefficient (ρ) may be regarded as a truer indication of the connection between the crop and each factor alone, inasmuch as, speaking approximately, we may say that the effect of the other factor is eliminated. It may be observed, moreover, that the relative influence of rainfall and temperature upon the crop is given by $\frac{\rho_{12}}{\rho_{13}}$; or, more accurately, this fraction measures the relative effect of changes equal in amount to their respective standard deviations in the rainfall and temperature. In discussing the figures in the tables I shall accordingly utilize the partial correlation coefficients rather than the others. Finally, I have worked out what Mr. Yule calls the coefficient of double correlation between the crop and rainfall and accumulated temperature above 42°,

$$R = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\cdot r_{23}\cdot r_{13}}{(1-r_{23}^2)}},$$

or as it may also be written,

$$R = \sqrt{1-(1-r_{12}^2)(1-\rho_{13}^2)},$$

a form which is quicker to calculate. This may be regarded

*Note on Estimating the Relative Influence of Two Variables upon a Third, Journal of the Royal Statistical Society, Vol. 69, pp. 197–200.

†Journal of the Royal Statistical Society, Vol. 70, pp. 5 and 6.

as a measure of the joint influence of the rainfall and the temperature upon the crop. For the sake of brevity, I shall speak of R as measuring the effect of the 'weather,' using this term in the strictly limited sense of consisting only of these two factors. . . .

"I propose to regard a coefficient between 0.3 and 0.5 as *suggestive* of dependence. Values below 0.3 I shall, as a rule, ignore, in the absence of any corroborative evidence. Perhaps I may remark that I believe that some statisticians would consider themselves justified in drawing deductions from lower coefficients than those I have adopted as my limits."*

Mr. Yule notes that the partial or net correlation coefficient retains three of the chief properties of the ordinary coefficients: "(1) it can only be zero if both net regressions are zero; (2) it is a symmetrical function of the variables (*i. e.*, $\rho_{12}=\rho_{21}$); (3) it cannot be greater than unity."†

The various illustrations which have been cited show the importance of questions of correlation in economics. The ordinary graphic method of measuring correlation is inadequate. The coefficient of correlation is simple and yet is sensitive to small changes. It has been used in many fields of statistics by Galton, Pearson, Yule, Hooker, Elderton and others. The experience of these writers warrants the adoption of the coefficient of correlation by economists as one of their standard averages.

*Journal of the Royal Statistical Society, Vol. 70, p. 5.

†*Ibid.*, Vol. 60, p. 833.